Anneleen Van Geystelen[1,2]*
Tom Wenseleers[2]
Ronny Decorte[1,3]
Maarten J. L. Caspers[1,4]
Maarten H. D. Larmuseau[1,3,4]*

[1]Laboratory of Forensic Genetics and Molecular Archaeology, UZ Leuven, Leuven, Belgium
[2]Laboratory of Socioecology and Social Evolution, Department of Biology, KU Leuven, Leuven, Belgium
[3]Biomedical Forensic Sciences, Department of Imaging & Pathology, KU Leuven, Leuven, Belgium
[4]Laboratory of Biodiversity and Evolutionary Genomics, Department of Biology, KU Leuven, Leuven, Belgium

Received September 17, 2013
Revised December 3, 2013
Accepted January 7, 2014

Research Article

In silico detection of phylogenetic informative Y-chromosomal single nucleotide polymorphisms from whole genome sequencing data

A state-of-the-art phylogeny of the human Y-chromosome is an essential tool for forensic genetics. The explosion of whole genome sequencing (WGS) data due to the rapid progress of next-generation sequencing facilities is useful to optimize and to increase the resolution of the phylogenetic Y-chromosomal tree. The most interesting Y-chromosomal variants to increase the phylogeny are SNPs (Y-SNPs) especially since the software to call them in WGS data and to genotype them in forensic assays has been optimized over the past years. The PENNY software presented here detects potentially phylogenetic interesting Y-SNPs in silico based on SNP calling data files and classifies them into different types according to their position in the currently used Y-chromosomal tree. The software utilized 790 available male WGS samples of which 172 had a high SNP calling quality. In total, 1269 Y-SNPs potentially capable of increasing the resolution of the Y-chromosomal phylogenetic tree were detected based on a first run with PENNY. Based on a test panel of 57 high-quality and 618 low-quality WGS samples, we could prove that these newly added Y-SNPs indeed increased the resolution of the phylogenetic Y-chromosomal analysis substantially. Finally, we performed a second run with PENNY whereby all samples including those of the test panel are used and this resulted in 509 additional phylogenetic promising Y-SNPs. By including these additional Y-SNPs, a final update of the present phylogenetic Y-chromosomal tree which is useful for forensic applications was generated. In order to find more convincing forensic interesting Y-SNPs with this PENNY software, the number of samples and variety of the haplogroups to which these samples belong needs to increase. The PENNY software (inclusive the user manual) is freely available on the website http://bio.kuleuven.be/eeb/lbeg/software.

Keywords:
Forensic genetics / SNP calling / Whole genome sequencing / Y-chromosome / Y-chromosomal SNP mutations
DOI 10.1002/elps.201300459

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

The human Y-chromosomal phylogenetic tree needs to be as accurate as possible since it has important applications in forensic sciences [1], evolutionary anthropology and population genetics [2–4]. Forensic scientists are taking advantage of the Y-chromosomal phylogenetic tree in their daily work, especially to assign geographical origins to specific lineages [5]. Y-chromosomal SNPs (Y-SNPs) have a great capacity to detect geographical origins as many lineages defined by Y-SNPs show a strong continent-specific [6, 7] and even intra-continent-specific distribution [8–10]. Their usefulness in forensic cases is illustrated by the fact that Y-SNP data are now also included in Y-chromosomal forensic databases, such as in the YHRD database [11]. Therefore,

Correspondence: Dr. Maarten H. D. Larmuseau, Biomedical Forensic Sciences, Department of Imaging & Pathology, Katholieke Universiteit Leuven, Kapucijnenvoer 33, Leuven B-3000, Belgium
E-mail: maarten.larmuseau@bio.kuleuven.be
Fax: +32-16324575

Abbreviations: ISOGG, International Society of Genetic Genealogy; **MCC**, Matthews correlation coefficient; **MRCA**, most recent common ancestor; **NGS**, next-generation sequencing; **WGS**, whole genome sequencing; **Y-SNP**, Y-chromosomal SNP

*Both authors contributed equally to this study.

several SNP arrays are currently used to genotype Y-SNPs for forensic applications [12, 13]. Hence, an up-to-date extended Y-chromosomal phylogeny is required using bi-allelic markers that are both unambiguous and non-recurrent and have high discrimination power [14].

The latest up-to-date Y-chromosomal tree for forensic applications was published by Van Geystelen et al. [14] and this tree was mainly based on the 'official' tree of Karafet et al. [7]. Several publications on lineage-specific updates of the latter tree have been made already [e.g. 15, 16]. All these updates are made based on sequencing limited parts of the Y-chromosome for a small group of individuals belonging to a limited set of sub-haplogroups. This means most likely that many interesting markers are still missing in the current phylogenetic tree. Over the last few years, rapid progress in next-generation sequencing (NGS) technologies has been made such that an increased number of human whole genome sequencing (WGS) data has become available [17, 18]. These WGS technologies cover the whole Y-chromosome, unlike the previous methods used to find novel Y-chromosomal markers for the phylogenetic tree. Therefore, it is expected that WGS data will become available in the future which will allow for optimizing and increasing the resolution of the current phylogenetic Y-chromosomal tree [19].

There are two methods to create an enhanced phylogenetic tree based on the Y-chromosomal markers found in the available WGS data. First, a new Y-chromosomal phylogenetic tree can be built after a tabula rasa of the present Y-chromosomal tree by using only variants from the available WGS male samples [20–22]. Recently developed phylogenetic methods make it relatively easy to obtain a maximum-likelihood tree [23, 24]. However, these trees do not provide links between the present and newly reported lineages. Such methods are therefore not useful for forensic sciences. Furthermore, these procedures require a large set of high-quality genome sequences to permit a full representation of all existing Y-chromosomal (sub-)haplogroups and geographical regions. However, such an ideal set is not yet available [14]. The second possible method consists of using the presently verified phylogeny and making changes to it: refining sub-haplogroups, resolving polytomies and adding extra markers that confirm existing sub-haplogroups. This method has an advantage in that it uses well-established data on the Y-chromosomal phylogeny and therefore provides a link between the present and newly reported lineages. Moreover, it is also possible to take into account the high number of false-positive and false-negative SNP calls which occur in any WGS method so far [25]. As long as the WGS methods are not fully optimized and as long as there is no ideal set of WGS samples representing the major Y-chromosomal haplogroups, it is assumed that the second method would be more effective in creating an enhanced phylogenetic tree [14].

Although several Y-chromosomal variants can be called from WGS data, Y-SNPs are preferred over small insertions and deletions (indels) and short tandem repeats (Y-STRs) since NGS technologies do not yet allow for accurate STR and indel calling [26]. Early analyses of WGS data provided thousands of novel Y-SNPs [27–29]. More recent meta-analysis of all available WGS samples has increased this list to tens of thousands of Y-SNP loci which have not yet been implemented in the latest Y-chromosomal phylogenetic tree [14, 19]. These new Y-SNPs may reveal new phylogenetic sub-haplogroups, unravel polytomies and detect potential mistakes in the current phylogenetic tree. Recently, a new software tool, namely AMY-tree, has been developed to determine Y-chromosome lineages and identify novel Y-SNPs using called Y-SNPs from WGS data [19]. However, there are as yet no software tools available to validate these newly observed Y-SNPs and to determine their phylogenetic value.

The development of software tools that can categorize these recently identified Y-SNPs is needed. The first reason for this requirement is again the high number of false-positive and false-negative SNP calls in any WGS analysis [25]. Incorrect Y-SNP calls can be made due to the highly repetitive character of the Y-chromosome which makes mapping to the reference genome very difficult [30]. As a 'wet-lab validation' of more than 100 000 Y-SNPs is not financially feasible, novel Y-SNPs should instead be validated in silico. The second reason is that both the phylogenetic position of each Y-SNP and its value for the tree should be determined based on a broad set of analysed WGS samples. Therefore, the aim of this study was to develop a user-friendly software tool which handles WGS data in order to be able to validate new Y-SNPs and to update the present Y-chromosomal phylogenetic tree. It was also the aim at this study to test the software on as many WGS data samples as possible and to present an updated Y-chromosomal tree which will be useful for all Y-chromosomal applications within forensic genetics.

2 Materials and methods

2.1 Datasets

Y-SNPs from as many multiple WGS experiments from different genomic projects were collected. The dataset consisted of 847 different samples of in total 745 different Y-chromosomes (Table 1). This was an expansion of the dataset used by Van Geystelen et al. [14] with 78 extra samples from the Personal Genome Project (PGP; www.personalgenomes.org) and 22 samples from Shen et al. [18]. As each of the different projects used different human genome versions, all called SNPs were converted to the Hg19 version. Some male genomes are analysed in several projects and two families are also present in the dataset (Supporting Information Table 1). The sub-haplogroup or phylogenetic lineage of each sample in the dataset was determined with the AMY-tree software [19]. The latest version 1.2 of the phylogenetic tree and the mutation conversion file were used for determining the sub-haplogroup [14]. The determined sub-haplogroups and the corresponding measure of Y-SNP call quality introduced in Van Geystelen et al. [14],

Table 1. Overview of the datasets used to run the developed PENNY software and to analyse its results

Project	Samples	Reference
1000 Genomes phase 1	526	The 1000 Genomes Project Consortium [17]
1000 Genomes pilot 1	77	Altshuler et al. [31]
Complete Genomics	35	Drmanac et al. [32]
Personal Genome Project	118	www.personalgenomes.org
Individual genome projects	45	Schuster et al. [28]; Peters et al. [33]; Rothberg et al. [34]; Chen et al. [35]; Tong et al. [29]; Wang et al. [36]; Ahn et al. [27]; Rasmussen et al. [37]; Wheeler et al. [38]; Pushkarev et al. [39]; Keller et al. [40]; Shen et al. [18]; www.everygenome.com; personal communication with Guy Froyen (VIB)
Singapore Sequencing Malay Project	46	Wong et al. [41]
Total samples	847	
Unique genomes	745	

that is, Matthews correlation coefficient (MCC), can be found in Supporting Information Table 2. AMY-tree also returned a list of all called Y-SNPs for each sample, which were not present in the current phylogenetic tree. In total, 847 WGS samples were collected for this study: 229 samples have a high-quality Y-SNP calling (MCC $\geq$ 0.95) while 618 have a low quality (MCC $<$ 0.95). A test panel was constructed to verify the quality of the PENNY results afterwards: 57 high-quality WGS samples which were uploaded between 10 April 2013 and 15 May 2013 on the PGP website were reserved for the validation phase.

In order to check if the called Y-SNPs were previously known, the latest update of dbSNP (version 138, 25 April 2013) was obtained from USCS (hgdownload.cse.ucsc.edu/goldenPath/hg19/database/snp137Common.txt.gz). The not yet peer-reviewed phylogenetic tree of International Society of Genetic Genealogy (ISOGG) contains more Y-SNPs than those in the current phylogenetic tree. Therefore, a list of all Y-SNPs present in the ISOGG phylogeny was obtained via their website (www.isogg.org, 20 November 2013).

2.2 PENNY software

The PENNY software (inclusive the user manual) is freely available on the website http://bio.kuleuven.be/eeb/lbeg/software. Based on the SNP calling input files, the PENNY software detects novel Y-SNPs which are potentially interesting for the Y-chromosome phylogeny by using a series of filters on all called Y-SNPs from WGS experiments (Fig. 1). The first filter is the selection of all Y-SNPs which occur in high-quality samples. The measure for Y-SNP calling quality and its threshold can be set by the user. We defined high quality as providing an MCC $\geq$ 0.95 which means that more than 97.5% of the negative and positive predictions are correct for such samples [14]. The second and third filters will remove all Y-SNPs which were found in only one sample and also any Y-SNP in non-unique regions of the Y-chromosome. A list of non-unique regions in the Y-chromosome (Supporting Information Table 3) was assembled based on information about pseudoautosomal, heterochromatic, X-transposed and ampliconic segments [30] of the male-specific part of the genome as reported by Wei et al. [20]. Next, information about paternal relatedness is taken into account in the fourth filter: if the new Y-SNP is present in more than 75% of all paternally related samples in the dataset then it passes the 'Relatedness' filter. However, if the Y-SNP is present in all paternally related samples and it is absent in all other samples, then this Y-SNP does not pass the filter and it is classified as a validated private Y-SNP. A threshold of 75% was chosen based on tests with the current dataset and based on the thresholds used in Van Geystelen et al. [14, 19] to define the sub-haplogroup of a specific sample taking the possibility of false negatives and false positives in WGS into account. Nevertheless, this value can be changed by the user. As mentioned before, the dataset also contains samples which come from the same

Figure 1. Schematic workflow of the PENNY software for the detection of Y-SNPs that are potentially interesting for the Y-chromosome phylogeny. For the detection of potentially interesting Y-SNPs, only the Y-SNPs of samples with a high SNP calling quality are used. These Y-SNPs need to pass another set of five filters before being classified as potentially interesting Y-SNPs. Validated private Y-SNPs can only be obtained when multiple samples from one individual or family is available.

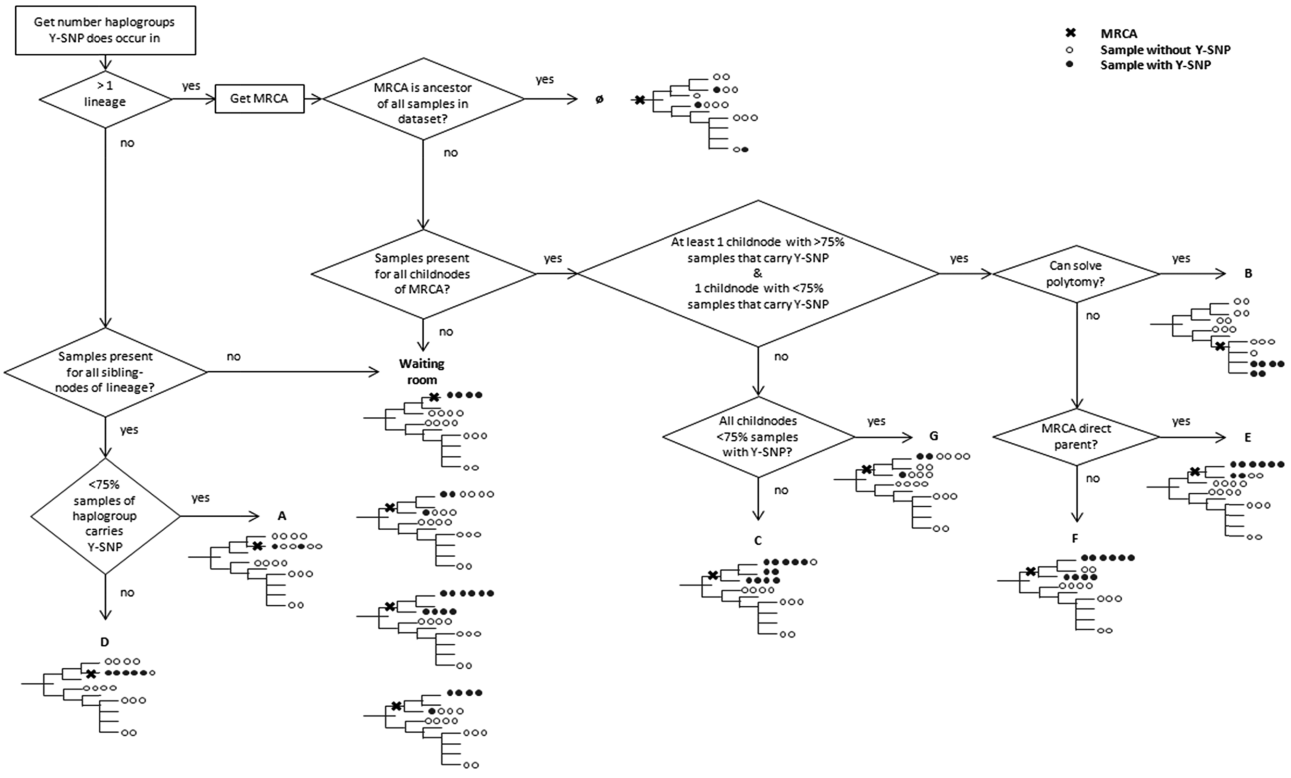

Figure 2. Workflow of the MRCA filter which divides the potentially interesting Y-SNPs into different types (A, B, C, D, E, F, G). For each type of potentially interesting Y-SNPs, a fictive example is given whereby ✖ stands for the most recent common ancestor (MRCA) of all samples that carry the Y-SNP, ○ stands for the samples that do not carry the Y-SNP and ● stands for the samples that do carry the Y-SNP.

DNA donor. These samples are of particular interest as they can point out possible false-positive Y-SNPs which should be avoided. The fifth filter uses this information. If a new Y-SNP is present in more than 75% of all samples of an identical genome then it passes the 'Identical' filter. However, if the Y-SNP is present in all samples of an identical genome and it is absent in all other samples, then this Y-SNP does not pass the filter, as it does not improve the phylogenetic analysis, and it is therefore a validated private Y-SNP. Again, a threshold of 75% was chosen based on tests with the current dataset. The last filter uses information about the most recent common ancestor (MRCA) of all samples that carry the Y-SNP. Figure 2 shows the workflow of this 'MRCA' filter.

The goal of the MRCA filter is to remove the phylogenetic uninformative Y-SNPs to indicate which Y-SNPs can become phylogenetic informative when more samples from other sub-haplogroups become available, and to divide the other Y-SNPs into different types (from A to G; Fig. 2). For both type A and type D, the Y-SNP needs to be carried by samples belonging to only one sub-haplogroup which is a leaf of the phylogenetic tree, and there must be samples present for all sibling nodes of the sub-haplogroup. If less than 75% of the samples belonging to that sub-haplogroup carry the Y-SNP, it is classed as type A else it is classed as type D. When there are no samples present in the dataset for any of the sibling nodes, the Y-SNP will be placed in the 'waiting room' as the exact phylogenetic position of the Y-SNP cannot (yet) be determined. This waiting room collection contains Y-SNPs which can become potentially interesting when the presence/absence of that Y-SNP in samples of the missing sub-haplogroups is added. If the Y-SNP occurs in multiple leaves, the MRCA of all Y-SNP carrying samples is determined. When that MRCA is the ancestor of all high-quality samples in the dataset, the Y-SNP will not pass the filter since it has no phylogenetic informative value. The next separator is whether or not there are samples present for each sub-haplogroup of the MRCA in the dataset. If not, the Y-SNP will be placed in the 'waiting room'; otherwise the Y-SNP is classified type B, C, E, F or G. If the Y-SNP is carried in more than 75% of the samples for each sub-haplogroup of the MRCA, it is classified as type C. If the Y-SNP on the other hand is carried in less than 75% of the samples for each sub-haplogroup of the MRCA, it is classified as type G. In all other cases some sub-haplogroups of the MRCA have ≥75% of their samples carrying the Y-SNP while other sub-haplogroups have <75% of their samples carrying the Y-SNP. For those Y-SNPs, the program checks if they can solve the polytomy of the MRCA. If they can, they are classified as type B, otherwise a distinction is made between the Y-SNPs. Either the MRCA is the direct parent of all samples that carry the Y-SNP (type E), or the MRCA is not the direct parent of all Y-SNP carrying samples (type F). When only one sample for any sub-haplogroup of the MRCA is available,

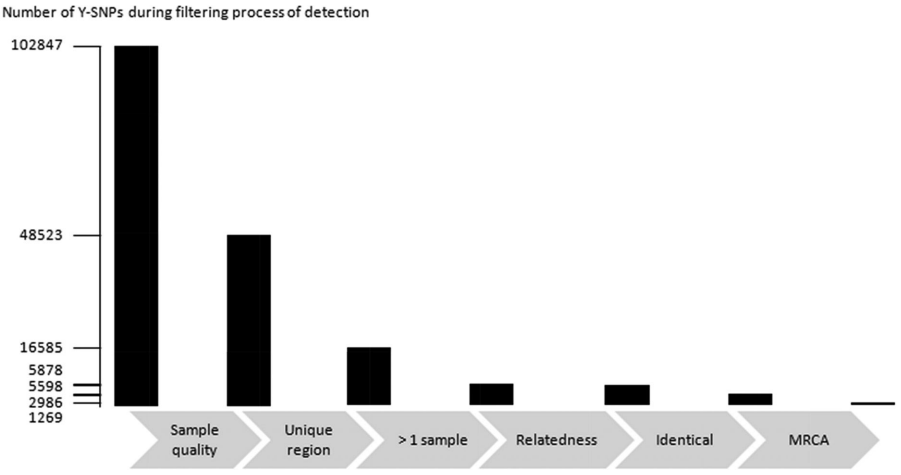

Figure 3. Evolution of the number of Y-SNPs during the filtering process for the detection of potentially interesting Y-SNPs.

then the type of the Y-SNP is indicated as (E) or (F) since there are not enough samples for this sub-haplogroup to be confident about the type. This 'MRCA' filter is also applied when the absence of a Y-SNP could be informative, that is, the mutant state is present in the reference genome and the ancestral state of the Y-SNP is called.

3 Results

3.1 Potentially phylogenetically interesting Y-SNPs

In total, 102 847 called Y-SNPs from the 790 WGS samples were used as the starting point for the detection of potentially interesting Y-SNPs for increasing the resolution of the Y-chromosomal tree. More than 50% of those Y-SNPs occur only in samples with an MCC $<$ 0.95 as Fig. 3 shows. Two-thirds of the remaining 48 523 Y-SNPs are located in non-unique regions of the Y-chromosome such that they are removed by the 'Unique regions' filter. Another two-thirds of those Y-SNPs are only carried by one sample and as such they are also removed. The 'Relatedness', 'Identical' and 'MRCA' filters reduce the number of potentially interesting Y-SNPs to 1269. All these potentially interesting Y-SNPs are listed in Supporting Information Table 4, and the names that were given to these novel Y-SNPs can be found in Supporting Information Table 5. Over 1600 Y-SNPs were placed in the 'waiting room' (Supporting Information Table 6) which means that currently they are not yet potentially of interest since there are samples missing for some sub-haplogroups of the MRCA. When the presence/absence of those Y-SNPs in samples of the missing sub-haplogroup(s) is included, they might become potentially interesting. Also 113 private Y-SNPs were validated based on information about their occurrence in multiple samples of the same genome (Supporting Information Table 7).

Figure 4 shows the number of Y-SNPs per type for all Y-SNPs and for a subset of non-equivalent Y-SNPs. When all Y-SNPs are considered more than half of the Y-SNPs belongs

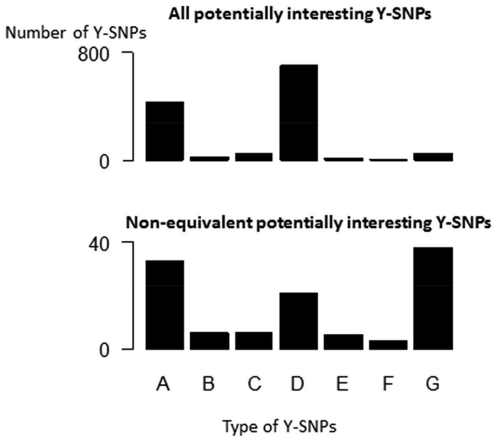

Figure 4. Number of potentially interesting Y-SNPs per type. Upper panel: all potentially interesting Y-SNPs. Lower panel: non-equivalent potentially interesting Y-SNPs.

to type D, another third belongs to type A and only very few Y-SNPs of types B, C, E, F and G are present in the dataset of potentially interesting Y-SNPs. As many Y-SNPs are equivalent, the distribution of the non-equivalent Y-SNPs per type looks different: the largest group is now type G followed by types A and D and the amount of Y-SNPs belonging to types

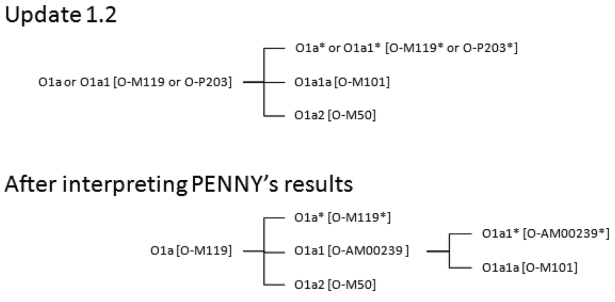

Figure 5. Phylogeny of O1a (O-M119) in update 1.2 (upper panel) and after interpretation of the results by the PENNY software (lower panel).

B, E and F are still relatively low. The relative high number of Y-SNPs belonging to type G indicates that those Y-SNPs which occur in different leaves of the tree with only a low frequency are probably false-positive or false-negative SNP calls in several samples. Examples of novel phylogenetically interesting Y-SNPs are given in Supporting Information Tables 8–14. Based on all detected informative Y-SNPs, an updated Y-chromosomal tree was composed (Supporting Information Tables 19 and 20). A manual control of all Y-SNPs was still required especially for SNPs of types F and G. The new updated tree is called version 2.0.

3.2 Validation of detected potentially interesting Y-SNPs

As validation of all potentially interesting Y-SNPs is impossible by laboratory analysis, one validation was performed by comparing the potentially interesting Y-SNPs with the Y-SNPs known by the ISOGG but which have not yet been included in the phylogenetic tree. As shown by Supporting Information Table 15, there were 27 Y-SNPs which were known by ISOGG, and for all except one, that is, Z9, the MRCA determined by PENNY corresponds with the lineage of ISOGG. Only seven Y-SNPs have less than ten equivalent Y-SNPs which means that these Y-SNPs support the validation.

A second validation method is based on the so called 'combined sub-haplogroups'. Van Geystelen et al. [19] had to combine several sub-haplogroups with their direct parent sub-haplogroup because they were poorly defined as a small insertion or deletion (indel), which are not yet called with sufficient quality in the WGS data, or contain a recurrent Y-SNP which negatively influences the assignment of the sample to a certain sub-haplogroup. However, these combined sub-haplogroups have the potential to permit the determination of potentially interesting Y-SNPs. Supporting Information Table 16 shows an example of how these specifically identified Y-SNPs can be placed into a combined haplogroup. For example, the combined sub-haplogroup O1a* or O1a1* (O-M119* or O-P203*) was created because the recurrent Y-SNP P203 was the only defining Y-SNP for sub-haplogroup O1a1* (O-P203*). Based on results from seven high-quality samples belonging to this combined sub-haplogroup PENNY indicated that 55 potentially interesting Y-SNPs could replace the recurrent Y-SNP P203 as a marker for this sub-haplogroup. Figure 5 shows the phylogenetic tree that was used in the analysis (update 1.2) and the phylogenetic tree after the interpretation of results using the PENNY software. Due to PENNY, it was possible to restore the phylogeny of O1a (O-M119) without the use of recurrent Y-SNPs. There are also combined sub-haplogroups created based on the presence of indels but their determination is more difficult since the calling quality of indels in NGS data is still insufficient. Supporting Information Table 17 shows an example of such a combined sub-haplogroup; although the presence of the indel could not be traced, samples of sub-haplogroup J2b2* or J2b2a (J-M241 or J-M99) can be divided in two different groups. Another combined sub-haplogroup based on an indel is sub-haplogroup O or O1 (O-M175 or O-MSY2.2). As shown in Supporting Information Table 18, the occurrence of the indel MSY2.2 could not be traced and the five type B Y-SNPs do not discriminate O1 from the other sub-haplogroups but instead these Y-SNPs seem to differentiate O3 from O1 and O2. In Supporting Information Fig. 3, the phylogenetic tree of O1 (O-M175) before and after the interpretation of PENNY's results is shown.

The most valuable validation method was the test set which consisted of 619 low-quality samples in the total dataset and 57 most recent PGP samples which were not used in the first PENNY analysis. All these samples were run with AMY-tree and the updated tree version 2.0 which was the result of the PENNY analysis. Based on this new tree, 94 samples out of the 619 low-quality samples (15%) were assigned to a phylogenetic deeper lineage or sub-haplogroup in comparison with the updated tree version 1.2. For several of these samples, a newly added Y-SNP was called revealing an assignment to a much phylogenetic deeper lineages, for example, from R1* (R-M173*) to R1b1b2a1a2d2 (R-AM00492). The average MCC of these 619 low-quality samples increased using the updated tree version 2.0 in comparison to the tree version 1.2. For the test set with 57 high-quality samples, 12 samples (21%) were assigned to a phylogenetic deeper lineage or sub-haplogroup in comparison with the updated tree version 1.2. The averaged MCC of these 57 high-quality samples increased, with updated tree 2.0 in comparison with tree 1.2. Moreover, the discrimination power increased because most of the 12 samples belonged to a newly discovered group which was not included with the other 45 samples in the test set based on the Y-SNPs of the updated tree version 1.2.

3.3 Updated tree

After the positive validation of PENNY which lead to version 2.0 of the Y-chromosomal phylogenetic tree, we performed a second run with PENNY including the 57 high-quality samples from the test panel. We also added all new Y-SNPs and evolutionary lineages which were published till 20 November 2013, including Rocca et al. [42], Mendez et al. [43], van Oven et al. [44], Poznik et al. [22], Francalacci et al. [21], Di Cristofaro et al. [45] to the tree. This second run resulted in 509 additional phylogenetic promising Y-SNPs. Finally, an updated Y-chromosomal tree, version 2.1, was composed with in total 582 evolutionary lineages and 2454 Y-SNPs (Supporting Information Tables 21 and 22).

4 Discussion

The results of this study show that the PENNY software was successful in increasing the resolution of the phylogenetic tree for forensic applications. Using 790 WGS samples in a

first run with the PENNY software, 1269 potentially informative Y-SNPs were validated in silico based on strict criteria. Another 1638 Y-SNPs were placed in the 'waiting room' since there were yet not enough samples in the dataset available to be sure about their specific phylogenetic position. Since the dataset also contained samples of paternally related males and samples which were derived from the same male but analysed in different projects, 113 Y-SNPs specifically for one family or individual, the so-called 'private' Y-SNPs, were validated in silico. In a second run with the PENNY software with in total 847 WGS samples, 509 additional phylogenetic promising Y-SNPs were detected. The strength of PENNY is the classification of the informative Y-SNPs into several types according to their position in the phylogeny, and their application towards improving the resolution of the tree (Fig. 2). PENNY indicates type A Y-SNPs which can refine existing end-lineages of the tree and therefore increase the discrimination power of samples. Another important type of Y-SNPs pointed out by PENNY is type B; these Y-SNPs solve the polytomies which are numerously present in the currently used phylogenetic tree. Also confirmation of the existing lineages is important; Y-SNPs which do confirm existing sub-haplogroups are classified as types C and D for internal nodes and end-leaves, respectively. Finally, types F and G of Y-SNPs are important for identifying any mistakes in the current Y-chromosomal tree. The value of the newly Y-SNPs in an updated tree and the higher discrimination power of this tree after the first run of PENNY was confirmed by a test set of 57 recent high-quality and 619 low-quality WGS samples.

The final result of this study was the improved phylogenetic Y-chromosomal tree (Supporting Information Tables 21 and 22). This updated phylogenetic tree has a much higher resolution due to the phylogenetic promising Y-SNPs which were pointed out by the two runs with the PENNY software. PENNY is responsible for (i) adding in total 25 evolutionary lineages to the phylogenetic tree, (ii) solving some polytomies of the previous phylogeny and (iii) increasing the number of Y-SNPs in the tree with 1394. By including new WGS data of high quality in the future, novel Y-SNPs will be identified and some of the Y-SNPs in the so-called 'waiting room' will be assigned to a Y-SNP type. Moreover, the validated 'private' Y-SNPs in this study may also be included in the phylogenetic tree in the future when unrelated WGS samples with these 'private' SNPs will also be available as they will show the relevance of these SNPs. Of course, the phylogenetic position of the Y-SNPs which are included in the tree in this study and which passed all filters based on the current dataset may change in the future or they may even disappear, as WGS improves the resolving power.

Although the criteria used to include Y-SNPs into the updated phylogenetic tree were quite strict, a relatively low number of potentially interesting Y-SNPs were found in silico by PENNY when considering that the dataset consisted of 172 high-quality genomes with an average of 1994 called Y-SNPs. The most likely explanation for this observation is that the backbone of the current phylogenetic tree is already well established and therefore only a limited number of potentially phylogenetically interesting Y-SNPs were found. With the exception of a few Y-SNPs classified in types F and G, there were no Y-SNPs which contradict the currently used Y-chromosomal phylogeny. This was also no surprise as Wei et al. [20] already confirmed the backbone of the tree using the tabula rasa method. Finally, several sub-haplogroups in the previous updated tree were a combination of two lineages as one was defined by an indel or a recurrent mutation which are not efficient for an AMY-tree analysis [19]. Due to the new Y-SNPs found by PENNY, it was possible to split many of those combined groups again. Therefore, this is again an indication that PENNY may find most of the relevant sub-haplogroups of the phylogeny and that most of the main sub-haplogroups are already known.

Another explanation for the low number of potentially interesting Y-SNPs is the relatively limited range of high-quality samples that was used in the PENNY analysis. Supporting Information Fig. 4 shows the distribution of the haplogroups to which the high-quality samples belong. Most samples belong to haplogroups R, O and I and therefore it is not surprising that these haplogroups represent the most non-equivalent potentially interesting Y-SNP, namely 38, 31 and 21%, respectively. Moreover, as most Y-chromosomal research is done in Eurasia where haplogroups R, O and I have the highest frequencies, it may be unsurprising that a relatively low number of novel potentially interesting Y-SNPs were found [10]. Therefore, an ideal dataset for PENNY would consist of multiple high-quality WGS samples from each sub-haplogroup. In this ideal setting, the exact phylogenetic position of each potentially interesting Y-SNP could be determined. Of course it will be very hard to create this ideal dataset since sequencing with a high coverage of the whole genome is expensive, particularly when at least two samples of each of the now 582 sub-haplogroups need to be sequenced. Another difficulty to create this ideal dataset is the fact that not each sub-haplogroup occurs with the same frequency in the world population. For some sub-haplogroups it will be harder to find multiple samples than for others [6,7]. Although the creation of such an ideal dataset seems far away, every day new WGS data becomes available.

To summarise, the PENNY software provides the opportunity to refine and extend the current Y-chromosomal phylogenetic tree for forensic applications based on in silico detection of potentially interesting Y-SNPs which were called in WGS data of male samples. The PENNY analysis revealed a new updated phylogenetic tree with much higher resolution and discrimination power as observed with a test set of high- and low-quality WGS samples. Therefore, a more up-to-date Y-chromosomal phylogenetic tree can be compiled for forensic applications. Although PENNY is a very useful program to find in silico novel potentially interesting SNPs for the Y-chromosome phylogeny, it has a major drawback, namely that it depends on high-quality sample data. This high-quality standard is required as the lineage of a sampled Y-chromosome in the present Y-chromosomal tree must be known with a high certainty in order to localize the phylogenetic position of new detected Y-SNPs. As the number of

high-quality sample data is still limited, the output of the software will be better when more sub-haplogroup diverse samples become available. Projects such as the 1000 Genomes Project [17, 31] which aim to sequence whole genomes of a large number of people to provide a comprehensive resource on human genetic variation in several populations around the world would be perfect to create an ideal dataset for PENNY. However, these samples of the 1000 Genomes are sequenced with a low coverage and therefore they do not pass our quality filter. Moreover, exome sequencing, which is at the moment very popular, is not an option since most phylogenetically interesting Y-SNPs are not located in the relatively few genes present on the Y-chromosome. Specific initiatives are therefore needed to provide an ideal dataset of samples to improve the Y-chromosomal tree.

The authors want to thank Mannis van Oven, Manfred Kayser, Nancy Vanderheyden, Jean-Jacques Cassiman, Filip Volckaert, Tom Havenith, Marie Boz, Hendrik Larmuseau and Lucrece Lernout for useful assistance and discussions. Thanks to Guy Froyen (VIB, KU Leuven), Richard Rocca (independent researcher), Cuiping Pan (Stanford University) and Andreas Keller (Saarland University) for providing us yet unpublished and published called SNPs of whole genome sequencing projects. Maarten H.D. Larmuseau is postdoctoral fellow of the FWO-Vlaanderen (Research Foundation Flanders). This study was funded by the KU Leuven BOF-Centre of Excellence Financing on 'Eco and socio-evolutionary dynamics' (Project number PF/2010/07) and on 'Centre for Archaeological Sciences 2 (CAS 2) – New methods for research in demography and interregional exchange'.

The authors have declared no conflict of interest.

5 References

[1] Kayser, M., in: Rapley, R. D. W. (Ed.), *Molecular Forensics*, John Wiley & Sons Ltd., Chichesters 2007, pp. 141–161.

[2] Jobling, M. A., Tyler-Smith, C., *Nat. Rev. Genet.* 2003, *4*, 598–612.

[3] Larmuseau, M. H. D., Vanoverbeke, J., Gielis, G., Vanderheyden, N., Larmuseau, H. F. M., Decorte, R., *Heredity* 2012, *109*, 90–95.

[4] Underhill, P. A., Kivisild, T., *Annu. Rev. Genet.* 2007, *41*, 539–564.

[5] Butler, J. M., in: Butler, J. M. (Ed.), *Advanced Topics in Forensic DNA Typing: Methodology*, Academic Press, London 2011, pp. 371–403.

[6] Chiaroni, J., Underhill, P. A., Cavalli-Sforza, L. L., *Proc. Natl. Acad. Sci. USA* 2009, *106*, 20174–20179.

[7] Karafet, T. M., Mendez, F. L., Meilerman, M. B., Underhill, P. A., Zegura, S. L., Hammer, M. F., *Genome Res.* 2008, *18*, 830–838.

[8] Larmuseau, M. H. D., Vanderheyden, N., Jacobs, M., Coomans, M., Larno, L., Decorte, R., *Forensic Sci. Int. Genet.* 2011, *5*, 95–99.

[9] Cruciani, F., Trombetta, B., Antonelli, C., Pascone, R., Valesini, G., Scalzi, V., Vona, G., Melegh, B., Zagradisnik, B., Assum, G., Efremov, G. D., Sellitto, D., Scozzari, R., *Forensic Sci. Int. Genet.* 2011, *5*, E49–E52.

[10] Larmuseau, M. H. D., Vanderheyden, N., Van Geystelen, A., van Oven, M., Kayser, M., Decorte, R., *Forensic Sci. Int. Genet.* In press.

[11] Willuweit, S., Roewer, L., *Forensic Sci. Int. Genet.* 2007, *1*, 83–87.

[12] van Oven, M., Toscani, K., van den Tempel, N., Ralf, A., Kayser, M., *Electrophoresis* 2013, *34*, 3029–3038.

[13] van Oven, M., Ralf, A., Kayser, M., *Int. J. Legal Med.* 2011, *125*, 879–885.

[14] Van Geystelen, A., Decorte, R., Larmuseau, M. H. D., *Forensic Sci. Int. Genet.* 2013, *7*, 573–580.

[15] Trombetta, B., Cruciani, F., Sellitto, D., Scozzari, R., *PLoS One* 2011, *6*, e16073.

[16] Yan, S., Wang, C. C., Li, H., Li, S. L., Jin, L., Genographic Consortium. *Eur. J. Hum. Genet.* 2011, *19*, 1013–1015.

[17] 1000 Genomes Project Consortium, *Nature* 2012, *491*, 56–65.

[18] Shen, H., Li, J., Zhang, J., Chao, X., Jiang, Y., Wu, Z., Zhao, F., Liao, L., Chen, J., Lin, Y., Tian, Q., Papasian, C. J., Deng, H.-W., *Plos One* 2013, *8*, e59494.

[19] Van Geystelen, A., Decorte, R., Larmuseau, M. H. D., *BMC Genomics* 2013, *14*, 101.

[20] Wei, W., Ayub, Q., Chen, Y., McCarthy, S., Hou, Y., Carbone, I., Xue, Y., Tyler-Smith, C., *Genome Res.* 2013, *23*, 388–395.

[21] Francalacci, P., Morelli, L., Angius, A., Berutti, R., Reinier, F., Atzeni, R., Pilu, R., Busonero, F., Maschio, A., Zara, I., Sanna, D., Useli, A., Urru, M. F., Marcelli, M., Cusano, R., Oppo, M., Zoledziewska, M., Pitzalis, M., Deidda, F., Porcu, E. et al., *Science* 2013, *341*, 565–569.

[22] Poznik, G. D., Henn, B. M., Yee, M. C., Sliwerska, E., Euskirchen, G. M., Lin, A. A., Snyder, M., Quintana-Murci, L., Kidd, J. M., Underhill, P. A., Bustamante, C. D., *Science* 2013, *341*, 562–565.

[23] Price, M. N., Dehal, P. S., Arkin, A. P., *Plos One* 2010, *5*, e9490.

[24] Price, M. N., Dehal, P. S., Arkin, A. P., *Mol. Biol. Evol.* 2009, *26*, 1641–1650.

[25] Liu, Q., Guo, Y., Li, J., Long, J. R., Zhang, B., Shyr, Y., *BMC Genomics* 2012, *13*.

[26] Neuman, J. A., Isakov, O., Showron, N., *Brief. Bioinform.* 2013, *14*, 46–55.

[27] Ahn, S. M., Kim, T. H., Lee, S., Kim, D., Ghang, H., Kim, D. S., Kim, B. C., Kim, S. Y., Kim, W. Y., Kim, C., Park, D., Lee, Y. S., Kim, S., Reja, R., Jho, S., Kim, C. G., Cha, J. Y., Kim, K. H., Lee, B., Bhak, J., Kim, S. J., *Genome Res.* 2009, *19*, 1622–1629.

[28] Schuster, S. C., Miller, W., Ratan, A., Tomsho, L. P., Giardine, B., Kasson, L. R., Harris, R. S., Petersen, D. C., Zhao, F. Q., Qi, J., Alkan, C., Kidd, J. M., Sun, Y. Z., Drautz, D. I., Bouffard, P., Muzny, D. M., Reid, J. G., Nazareth, L. V., Wang, Q. Y., Burhans, R. et al., *Nature* 2010, *463*, 943–947.

[29] Tong, P., Prendergast, J. G. D., Lohan, A. J., Farrington, S. M., Cronin, S., Friel, N., Bradley, D. G., Hardiman, O.,

Evans, A., Wilson, J. F., Loftus, B., *Genome Biol.* 2010, *11*, R91.

[30] Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., Chinwalla, A., Delehaunty, A., Delehaunty, K., Du, H., Fewell, G., Fulton, L., Fulton, R., Graves, T., Hou, S. F., Latrielle, P. et al., *Nature* 2003, *423*, 825–837.

[31] Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De la Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., Nickerson, D., Peltonen, L. et al., *Nature* 2010, *467*, 1061–1073.

[32] Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K. P., Baccash, J., Borcherding, A. P., Brownley, A. et al., *Science* 2010, *327*, 78–81.

[33] Peters, B. A., Kermani, B. G., Sparks, A. B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y. T., Haas, J., Robasky, K., Zaranek, A. W., Lee, J. H., Ball, M. P., Peterson, J. E., Perazich, H., Yeung, G., Liu, J., Chen, L. S., Kennemer, M. I. et al.,*Nature* 2012, *487*, 190–195.

[34] Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A. et al., *Nature* 2011, *475*, 348–352.

[35] Chen, R., Mias, G. I., Li-Pook-Than, J., Jiang, L. H., Lam, H. Y. K., Chen, R., Miriami, E., Karczewski, K. J., Hariharan, M., Dewey, F. E., Cheng, Y., Clark, M. J., Im, H., Habegger, L., Balasubramanian, S., O'Huallachain, M., Dudley, J. T., Hillenmeyer, S., Haraksingh, R., Sharon, D. et al., *Cell* 2012, *148*, 1293–1307.

[36] Wang, J., Wang, W., Li, R. Q., Li, Y. R., Tian, G., Goodman, L., Fan, W., Zhang, J. Q., Li, J., Zhang, J. B., Guo, Y. R., Feng, B. X., Li, H., Lu, Y., Fang, X. D., Liang, H. Q., Du, Z. L., Li, D., Zhao, Y. Q., Hu, Y. J. et al., *Nature* 2008, *456*, U60–U61.

[37] Rasmussen, M., Li, Y. R., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., Metspalu, M., Metspalu, E., Kivisild, T., Gupta, R., Bertalan, M., Nielsen, K., Gilbert, M. T. P., Wang, Y., Raghavan, M., Campos, P. F., Kamp, H. M., Wilson, A. S., Gledhill, A., Tridico, S. et al., *Nature* 2010, *463*, 757–762.

[38] Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y. J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X. Z., Liu, Y., Yuan, Y. et al., *Nature* 2008, *452*, U872–U875.

[39] Pushkarev, D., Neff, N. F., Quake, S. R., *Nat. Biotechnol.* 2009, *27*, 847–850.

[40] Keller, A., Graefen, A., Ball, M., Matzas, M., Boisguerin, V., Maixner, F., Leidinger, P., Backes, C., Khairat, R., Forster, M., Stade, B., Franke, A., Mayer, J., Spangler, J., McLaughlin, S., Shah, M., Lee, C., Harkins, T. T., Sartori, A., Moreno-Estrada, A. et al., *Nat. Commun.* 2012, *3*, 698.

[41] Wong, L.-P., Ong, R. T.-H., Poh, W.-T., Liu, X., Chen, P., Li, R. Q., Lam, K. K.-Y., Pillai, N. E., Sim, K.-S., Xu, H., Sim, N.-L., Teo, S.-M., Foo, J.-N., Tan, L. W.-L., Lim, Y., Koo, S.-H., Gan, L. S.-H., Cheng, C.-Y., Wee, S., Yap, E. P.-H. et al., *Am. J. Hum. Genet.* 2013, *92*, 1–15.

[42] Rocca, R. A., Magoon, G., Reynolds, D. F., Krahn, T., Tilroe, V. O., Boots, P. M. O., Grierson, A. J., *Plos One* 2012, *7*, e41634.

[43] Mendez, F. L., Krahn, T., Schrack, B., Krahn, A. M., Veeramah, K. R., Woerner, A. E., Fomine, F. L. M., Bradman, N., Thomas, M. G., Karafet, T. M., Hammer, M. F., *Am. J. Hum. Genet.* 2013, *92*, 637–637.

[44] van Oven, M., Van Geystelen, A., Kayser, M., Decorte, R., Larmuseau, M. H. D., *Hum. Mutat.* 2014, *35*, 187–191.

[45] Di Cristofaro, J., Pennarun, E., Mazières, S., Myres, N. M., Lin, A. A., Temori, S. A., Metspalu, M., Metspalu, E., Witzel, M., King, R. J., Underhill, P., Villems, R., Chiaroni, J., *Plos One* 2013, *8*, e76748.

